Detecting Family Resemblance: Automated Genre Classification

Yunhyong Kim and Seamus Ross

Digital Curation Centre (DCC) & Humanities Advanced Technology Information Institute (HATII), University of Glasgow, Glasgow, UK
Email: {y.kim, s.ross}@hatii.arts.gla.ac.uk

ABSTRACT

This paper presents results in automated genre classification of digital documents in PDF format. It describes genre classification as an important ingredient in contextualising scientific data and in retrieving targetted material for improving research. The current paper compares the role of visual layout, stylistic features and language model features in clustering documents and presents results in retrieving five selected genres (Scientific Article, Thesis, Periodicals, Business Report, and Form) from a pool of materials populated with documents of the nineteen most popular genres found in our experimental data set.

keywords: automated genre classification, metadata, scientific information, information management, information extraction

1 INTRODUCTION

Scientific information is currently being created at an exponential rate and in many different forms (e.g. formally as scientific papers, raw data, as laboratory notes, or technical reports, and informally as emails, or letters). Even when a document does not give a direct description of scientific data or result it may provide the context essential for interpreting the scientific information: the context for understanding databases can often be found within scientific papers, and, in turn, the context for the results decribed in scientific papers can be found in informal discussions and logs in the form of emails, letters, laboratory notes or technical reports. It is only possible to keep track of these different information sources and relationships between data with metadata describing the content of the object. Manually acquiring such metadata is labour intensive and consequently expensive. The research in this paper reflects a motivation to automate the extraction of metadata from digital documents. Previous work exists on the extraction of descriptive metadata extraction within specific domains or genres (e.g. MetadataExtractor, DC-dot, Automatic Metadata Generation, Thoma (2001), Giuffrida, Shek & Yang (2000), Han, Giles, Manavoglu, Zha, Zhang & Fox (2000), Bekkerman, McCallum & Huang (2004), Ke, Bowerman & Oakes (2006), Sebastiani (2002) and Witte, Krestel & Bergler (2005)). However, a general tool has yet to be developed to extract metadata from documents of varied forms and subjects. This paper is a continuation of the work in Kim & Ross (2006a), Kim & Ross (2006b) and Kim & Ross (2006c) to develop genre classification; the automatic detection of document types, followed by deeper metadata extraction from single document types using domain-specific methods, as a means of creating an over-arching tool which can extract metadata across many domains at different semantic levels. The focus on genre classification acknowledges the fact that different communities focus on scientific materials in different genres: genre classification will support automating the identification, selection, and acquisition of materials in keeping with the goals of different scientific communities.

As we discussed earlier (Kim & Ross (2006a)) there is, however, a lack of consensus with regard to the definition of genre: Biber's analysis (Biber (1995)) of document genres employed five dimensions (information, narration, elaboration, persuasion, abstraction) to characterise text while others (Karlgren & Cutting (1994), Boese (2005)) examined popularly recognised genre classes such as FAQ, Job Description, Editorial or Reportage. There were attempts (Kessler, Nunberg & Schuetze (1997), Finn & Kushmerick (2006)) to automatically detect limited facets (narravative, objectivity, intended level of audience, positive or negative opinion) and an attempt (Bagdanov & Worring (2001)) to distinguish specific journals and brochures from one another. Others (Rauber & Mueller-Koegler (2001), Barbu, Heroux, Adam & Turpin (2005)) have clustered documents into similar feature groups without delving into genre facets or classes. An overview of the various efforts in genre analysis can be found in a technical report by Santini (2004a) and Santini (2004b) and a report

on metadata extraction by Dobreva, Kim & Ross[1] . The definition of genre adopted by these researchers all rely on a combination of two notions: one of structure and one of function. *Structure* is defined by factors which are reflected in the visual layout of the document while *function* is defined by the intended purpose of the document. The two notions are closely related: the structure of the document is formed to optimise the function of the document within an environment, such as within the context of the community or event, in which the document is created.

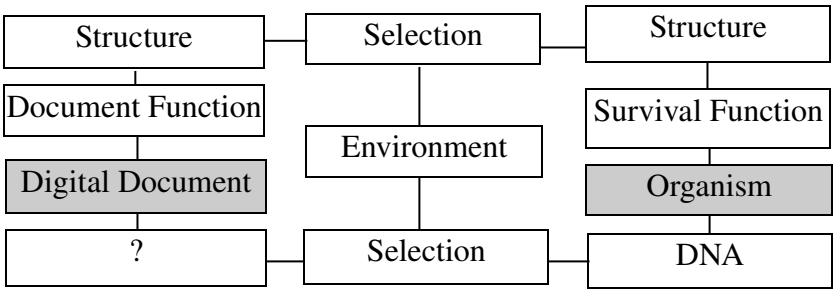

Figure 1: Evolution: document as a dynamic entity

The situation can be compared to the process of natural selection in the theory of evolution (Figure 1). There are basic functions required for an organism to survive within the environment. The structures with properties to optimise the survival functions are most likely to survive. These structures are a result of the expression of the genes in the DNA sequence which represents an organism: the entities within a species with genes which accompany the best structural properties will prosper. The question lies in determining the representation of a document which constitutes a DNA sequence, and further, to determine how *genetic information* of documents is encoded in the representation to characterise the document type. As the question mark in Figure 1 indicates, we feel that there is yet no proper understanding of what the DNA sequence of a document should be.

In this paper, we propose five types of features as a candidate subsequence of DNA for documents: image features, syntactic features, stylistic features, semantic structure, and domain knowledge features. We aim to eventually model the five types of features, and additional features, if necessary, to predict the genres of PDF documents from a working schema of seventy genres which was described in Kim & Ross (2006a). In this paper, we will examine the nineteen most prolific PDF (Adobe Acrobat PDF) genres (Table 1) found in our data set. The 570 PDF files in our data set were collected over the Internet by choosing a random word from a dictionary and retrieving a random PDF from the list returned by a search engine (details in Kim & Ross (2006a)). These genres are expected to represent the most popular PDF genres in the public domain. The main reason for choosing to work with PDF files is the fact that it is a portable widely used format for archived digital materials in scientific repositories.

Table 1: Reduced Scope of Genres

Groups	**Genres**
Book	Academic book, Fiction, Other book
Article	Scientific research article, Other research article, Magazine article
Serial	Peridocials (Newpaper, Magazine), Newsletter
Treatise	Thesis, Business/Operational report, Technical report
Information Structure	List, Form

1Funded by DELOS[11]. Expected to be available December, 2006

Evidential Document	Minutes
Other Functional Document	Guideline, Job/Course/Project Description, Product/Application Description, Fact sheet, Slides

The experiments in this paper are motivated by the conviction that a comparison of classifiers built separately on different types of features to analyse which genre is best distinguished by which classifier is an crucial step in the construction of a general classifier. Once the genres are better ubderstood in terms of their feature strengths, classifiers can be combined in an intelligent way to create a final prototype. Just as statistically modeling the entire DNA sequences for several species does not produce a high-level performance in the automatic detection of family resemblance, it does not seem reasonable to believe that statistically bundling up all these features would result in an optimal automatic classification system of document types. If all features are processed in one classifier, the statistical model can be misled by non-distinguishing features. If we were to train on sufficient data, this would not be a problem; the non-distinguishing features will be filtered out as noise. It is, however, very difficult to have *sufficient data* when constructing a tool which is intended to have dynamic and domain-independent properties. In Kim (2004) and Kim & Webber (2006), the CANDC part-of-speech tagger (Curran & Clark (2003)), reputed to have performed well elsewhere, was employed to tag words in Astronomy research articles. In Astronomy there is frequent usage of the term *He* to refer to the chemical element Helium. The tagger, which was trained on the *Wall Street Journal* articles, tagged *He* to be a pronoun for all instances, propagating further errors on subsequent words. Separating features into smaller groups will minimise the impact of such artefacts, by trying to exclude the noise from the start, making the most of the differing feature strengths for each genre type.

This paper, along with Kim & Ross (2006a), Kim & Ross (2006b) and Kim & Ross (2006c) also emphasises that the bottom-up approach of starting from genre-specific extraction may result in several tools which are overly dependent on the structures of the documents within a specific domain, with no obvious means of interoperability. The top-down approach of creating a tool which stretches across genres, to be refined further within the domain, will enable us to avoid this problem.

2 CLASSIFIERS

The experiments described in this paper involve the use of three classifiers:

Image classifier:this classifier depends on features extracted from the PDF document when handled as an image. It uses the module *pdftoppm* from XPDF (Noonberg (2006)) to extract the first page of the document as an image. The resulting image is divided into a sixty-six by sixty-six grid[2] . Then Python's Image Library (PIL) is employed to extract pixel values in each region. Each region is given a value of 0 or 1 depending on whether there is more than one pixel darker than a specified value of 245. The result is modeled using Naïve Bayes as implemented by the Weka (Witten & Frank (2005)) machine learning toolkit.

Language model classifier:this classifier depends on an *N*-gram model on the level of words, Part-of-Speech tags and Partial Parsing tags. N-gram models look at the possibility of unit $W(N)$ coming after a string of units $W(1),W(2),...,w(N-1)$. A popular model is the case when N=3. In the experiments of the current paper, we are only working with the model where W(i) are words. This can be modeled for other syntactic units or semantic units to capture the patterns of higher level structures. This has been modelled by the BOW toolkit developed by Andrew McCallum (1998). We used the default Naïve Bayes model.

Stylo-metric classifier:this classifier looks at the frequency of selected words, number of font changes, the difference between the largest font size and smallest font size, length of the document , average length of words, and number of words in the front page of the document. The font information was extracted on the level of words using a modified version of PDFTOHTML, developed as part of the DELOS Digital Preservation Cluster

2 The choice of the dimension reflects the fact that it seemed to produce the best results at the time but further analysis may be necessary.

investigations at Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI), University of Cologne. The modified version converts a PDF document to an XML file with the font size and style information for each word in the document. A word list was automatically constructed containing all words which appear in more than half of the files in any one genre. For each file, the frequency of each word was recorded as a vector then augmented by length and font information. The result was modeled using Naïve Bayes in the Weka (Witten & Frank (2005)) machine learning toolkit.

We have chosen to use Naïve Bayes method for all the classifiers. This is because the experiments are intended to compare different feature sets given similar conditions. Further analysis will be required to make a comparison of the effects of different statistical models. Naive Bayes was chosen for the experiments in this paper because it gave the best overall results for all the classifiers.

The view in this paper is that the image along with the stylistic features will capture the structural elements of genres while the language model combined with the stylistic and semantic features will help to separate documents of distinct functional categories.

Involving the image of a document in the process may enable genre classification of documents without violating pasword protection or copyright, will maximise the viability of a language independent tool and free the process from being solely dependent on text processing tools with encoding requirements and problems relating to special characters[3]. It also makes part of the process immediately applicable to paper documents digitally imaged (i.e. scanned).

3 EXPERIMENTS

Two main experiments are described in this paper:

Clustering experiment:in this experiment we compared the cluster resolution for three sets of features: the image features, the stylo-metric features, and the case where the two features were combined. We grouped the data in nineteen genres into two clusters using the Weka Machine Learning Toolkit's (Witten & Frank (2005)) Estimation-Maximisation algorithm. The purpose was to see how well the files in each genre group into one cluster. The result is expressed in terms of the percentage of files within each genre which have been grouped into one cluster.

A Comparison of Binary Predictions: In this experiment we will compare the 10-fold cross validation test of the three classifiers described in Section 2 in the following binary predictions.

• **Periodicals versus Other Genre:** in Kim & Ross (2006c) we presented the results of distinguishing Periodicals from Thesis and Periodicals from Non-periodicals consisting of Thesis, Business Reports, Minutes, Academic Book, and Fictional Book. In this experiment, we have expanded the class of non-periodicals to Other Genre to include all of the genres in Table 1 apart from periodicals.

• **Scientific Research Article versus Other Genre:** in this experiment, we test the binary classification of genres in Table 1 into Scientific Research Article and Other Genre.

• **Thesis versus Other Genres :** in this experiment, we test the binary classification of genres in Table 1 into Thesis and Other Genre.

We also compared the image classifier to the stylo-metric classifier in the following binary predictions:

• **Business Report versus Other Genres:** in this experiment, we test the binary classification of genres in Table 1 into Thesis and Other Genre.

• **Forms versus Other Genres :** in this experiment, we test the binary classification of genres in Table 1 into Thesis and Other Genre.

Finally, we also present the results for detecting Forms, using the stylo-metric classifier, within a group of files populated with a smaller variety of genres including Fact Sheet, Forms, Instruction/Guideline, Job/Course/Project Description, Minutes, Newsletter, Scientific Reserch Article and Other research Article.

---

3 *pdftohtml* failed to extract information from seventeen percent of the documents. The image processing did not fail on any documents.

4 RESULTS

Table 2 shows the results of the clustering experiment. The percentages for the visual clusters are different from the previously reported clusters (Kim & Ross (2006a)) because we have excluded the files for which *pdftohtml* failed to extract the correct information. The results of this experiment encourages two conclusions:

1. The genres for which image features fail to show a sharp clustering tendency (> 80) tend to be the genres for which stylo-metric features cluster very well and vice versa. In the case of only three genres (Forms, Job Description, and Product Description) this does not hold. For instance, note that, stylistic features only group 62.5 percent of files in the class Periodicals into one cluster whereas the visual features group one hundred percent of the files into one cluster.

2. Clustering based on a combination of both types of features at the same time does not improve upon clustering based on features of one type. In fact the combined system, as the third column of Table 2 shows, results in blurring the division for most genres. And, in the the case of Forms and Scientific Articles, the combined features show poorer clustering results than either of the clustering results in column one and two.

Table 2: A Comparison of Visual and Stylo-metric Clusters (percentage of files in one cluster)

Groups	***Genres***	***visual***	***stylistic***	***combi***
Book	Academic	100%	60%	100%
	Fiction	92.8%	83.3%	75%
	Other Book	70.6%	82.4%	70.6%
Article	Sci. Research Article	76%	92%	64%
	Other Research Article	94.7%	73.7%	84.2%
	Magazine Article	61.5%	84.6%	61.5%
Serial	Periodicals (Newspapers, Magazine)	100%	63%	88%
	Newsletter	54.2%	83.3%	58.3%
Treatise	Thesis	100%	90%	90%
	Business/Operational Report	81.8%	90.9%	81.8%
	Technical Report	88.9%	72.2%	83.3%
Information Structure	List	71%	86%	71%
	Forms	61.5%	69.2%	53.8%
Evidential Document	Minutes	100%	77%	100%
Other Functional Documents	Instruction/Guideline	95%	79%	90%
	Job/Course/Project Description	73.3%	50%	73.3%
	Product/Application Description	62.5%	66.7%	56.3%
	Factsheet	72.4%	85.7%	64.3%
	Slides	61.5%	91.7%	61.5%

The results described in Tables 3, 4, 5, 6, 7 and 8 use three standard indices in classification tasks: accuracy, precision and recall. Let N be the total number of documents in the data, NC the number of documents in the data set which are in class C, T the total number of correctly labelled documents in the data set independent of the class, TC the number of true positives for class C, and FC the number of false positives for class C. Accuracy is defined to be T/N; precision and recall for class C is defined to be $TC/(TC+FC)$ and TC/NC, respectively.

In [22] the image classifier was compared against the stylo-metric classifier to show that the image classifier performed much better in distinguishing Periodicals from the group of Non-periodicals (consisting of Thesis, Business Report, Minutes, Fictional Book and Academic Book). In Table3, we have presented the results of a 10-fold cross validation test using the image, stylo-metric and language model classifiers (in respective order from top to bottom) on detecting Periodicals. If we put only the overall accuracy into consideration, the language

model seems to be the best classifier, but, the recall rates show that the language model classifier failed to label even a single periodical correctly. The recall rate for the stylo-metric classifier fares better but it is nowhere near that of the image classifier. The language model classifier and the stylo-metric classifier are labelling every item or almost every item as Other Genre, thereby making no distinction between Periodicals and Non-periodicals; the higher percentage of Other Genre in experimental data set ensures a high overall accuracy rate for a blind system which labels all files as Other Genre. The precision of the image classifier for Periodicals may seem rather low, but it is important to keep in mind that the number of Periodicals in the whole dataset is only five percent of the number of files in Other Genre; only a small percentage of mislabelling in the Other Genre would result in a sizable drop of precision for Periodicals. The overall accuracy of the image classifier shows a slight decrease when compared to the previous experiment (Kim & Ross(2006c)) when the variety of non-periodical genres was smaller but the results in the two cases seem comparable. The result suggests the image features as a distinguishing factor for Periodicals.

Table 3: Distinguishing Periodicals from Other Genre using image (top) stylo-metric (middle) and language model (bottom)

10 fold Cross Validation with the Image classifier, Overall accuracy: 88.6%

Genre	Precision(%)	Recall(%)
Periodicals (16 items)	29.8	87.5
Other Genre (291 items)	99.2	88.7

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 88.52%

Genre	Precision(%)	Recall(%)
Periodicals (16 items)	14.8	25
Other Genre (291 items)	92	93.8

10 fold Cross Validation with the language model classifier, Overall accuracy: 94.79%

Genre	Precision(%)	Recall(%)
Periodicals (16 items)	0	0
Other Genre (291 items)	96.9	100

Table 4 shows the 10-fold cross validation results of the three classifiers on distinguishing Scientific Research Articles from Other Genre. As in the case of Periodicals, the overall accuracy is highest for the language model classifier but careful examination shows that the language classifier labels only fifteen percent of the Scientific Research Articles correctly. The best recall rate for Scientific Research Articles is found in the image classifier. However, the recall rate for Scientific Research Articles using the image classifier is only four percent better than the stylo-metric classifier while the precision falls more than twenty eight percent below that of the stylo-metric classifier. The overall performance seems to be best with the stylo-metric classifier.

Table 4: Distinguishing Scientific Research Articles from Other Genre using image (top), stylo-metrics (middle) and language model (bottom)

10 fold Cross Validation with the Image classifier, Overall accuracy: 73.94 %

Genres	Precision(%)	Recall(%)
Scientific Research Articles (25 items)	21.11	80
OtherGenre (280 items)	97.6	73.4

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 91.80 %

Genres	Precision(%)	Recall(%)
Scientific Research Articles (25 items)	50	76
OtherGenre (280 items)	97.8	93.2

10 fold Cross Validation with the language model classifier, Overall accuracy: 94.68 %

Genres	Precision(%)	Recall(%)
Scientific Research Articles (25 items)	100	15
OtherGenre (280 items)	94.6	100

Similar analysis of results for Thesis (Tables 5) seem to indicate the image features as the best distinguishing factor among the three types of feature sets for the category Thesis.

Table 5: Distinguishing Thesis from Other Genre using image(top) stylo-metrics (middle) and language model (bottom)

10 fold Cross Validation with the Image classifier, Overall accuracy: 82.74 %

Genres	Precision(%)	Recall(%)
Thesis (10 items)	13.6	80
OtherGenre (280 items)	99.2	90.3

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 75.40 %

Genres	Precision(%)	Recall(%)
Thesis (10 items)	7	60
OtherGenre (280 items)	98.2	75.9

10 fold Cross Validation with the language model classifier, Overall accuracy: 93.87 %

Genres	Precision(%)	Recall(%)
Thesis (10 items)	40	17.4
OtherGenre (280 items)	98	95.67

For Business Report and Forms (Tables 6, 7) seem to indicate stylo-metric features as a better distinguishing factor for detecting these genres. In fact, Table 8 shows that, in a 10-fold cross validation classification experiment using the stylo-metric classifier of files belonging to one of the classes Fact Sheet, Forms, Instruction Guidelines, Job/Course/Project Description, Minutes, Newsletter, Scientific Research Article and Other Research Article, Forms achieves a Recall of 92.3% and a Precision of 70.6%.

Table 6: Distinguishing Business Report from Other Genre using image(top) and stylo-metrics (bottom)

10 fold Cross Validation with the Image classifier, Overall accuracy: 60.72 %

Genres	Precision(%)	Recall(%)

Business Report (10 items)	5.6	63.6
OtherGenre (280 items)	97.8	60.1

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 72.79 %

Genres	Precision(%)	Recall(%)
Business Report (10 items)	9.1	72.7
OtherGenre (280 items)	98.6	72.8

Table 7: Distinguishing Forms from Other Genre using image(top) stylo-metric (bottom)

10 fold Cross Validation with the Image classifier, Overall accuracy: 76.55 %

Genres	Precision(%)	Recall(%)
Forms (10 items)	7.2	38.5
OtherGenre (280 items)	96.6	78.2

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 71.48 %

Genres	Precision(%)	Recall(%)
Forms (10 items)	10.6	76.9
OtherGenre (280 items)	98.6	71.2

Table 8: Classification of files into eight classes

10 fold Cross Validation with the stylo-metric classifier, Overall accuracy: 88.11 %

Genres	Precision(%)	Recall(%)
Fact Sheet (14 items)	27.3	21.4
Forms (10 items)	41.7	76.9
Instruction (20 items)	53.8	35
Job/Course/Proj.Desc. (15 items)	22.2	26.7
Minutes (13 items)	75	69.2
Newsletter (24 items)	48.4	62.5
Sci.Res.Article (25 items)	62.1	72
Other Res.Article (19 items)	50	31.6

Finally, Table 9 shows the results of using the language model classifier to predict the classes Thesis, Fictional Book, Academic Book, Minutes, and Business Report. The results show that, when the variety of genres is limited, the language model classifier can do very well on Business Report and Fictional Book and Minutes, which leads us to suggest, along with the results in Tables 6 and 8, that combining the stylo-metric classifier with the language model classifier may result in a classifier able to detect these classes.

Table 9: Classifying five types of genres using language model

10 fold Cross Validation with language model classifier, Overall accuracy on 5 classes: 82.4%

Genre	Precision(%)	Recall(%)
Academic Book (5 items)	42.9	60
Business Report (11 items)	90	81.8
Fictional Book (14 items)	100	100
Minutes (13 items)	86.7	92.9
Thesis (10 items)	75	60

5 CONCLUSION

The results in Kim & Ross (2006a), Kim & Ross (2006c), and this paper indicate the promise of focusing similar feature types to extract documents of selected genre classes. The results in Table 2 illustrate definite divisions between genres which have strong image features and genres that have strong stylistic features. They also show that combining features into one boiling pot does not necessarily improve the clusterer. The results in Tables 3 and 5 indicate that the classes Periodicals and Thesis have more clearly distinguishing image features than stylo-metric features or language model features. The figures in Table 4 shows that the tendency is less clear in the case of the class Scientific Research Article. Forms and Business Reports seem to have stronger stylo-metric features (Tables 7, 8 and 6) and extracting some genres may become more viable by incorporating the language model classifier (Table 9). There is still more work to do before we arrive at a tool which can classify files into nineteen genres or more. However, the results in this paper indicate that, by targetting selected genres and determining their strong feature types, we can narrow down the search for files in the selected genres within a pool of other documents; perhaps a more promising approach than to create a general tool which classifies files within a range of genre classes that keep changing in size and variety.

Further improvement can also be envisioned by integrating more classifiers into the decision making process. In Kim & Ross (2006a) we suggested the following classifiers: We could consider an **Extended image classifier** which looks at more than the first page of the document. This would involve decisions on the optimal number of pages to be used and the best way to statistically combine the information from different pages. The **Language model classifier** could be built on the level of part-of-speech tags (tags which denote whether a word is a verb, noun or preposition) or partial chunk tags (tags indicating noun phrases, verb phrases or prepositional phrases). A **Semantic classifier** could be employed to model subjective or objective noun phrases (e.g. using Riloff, Wiebe & Wilson (2003)) and latent semantic analysis. A **Contextual Classifier** built on source information of the document such as the name of the journal or address of the web page, anchor text or domain subject information, along with administrative organisational context when available, would be a useful addition.

There are two obvious ways of gauging the performance of a classifier: comparing against human performance and measuring the stability of the performance as you transfer it across domains. We are undertaking an experiment to examine human performance. A significant amount of disagreement is expected in labelling

genres even between human labellers; we intend to cross check the labelled data in two ways:

1. **Document Retrieval Exercise (DRE):** we are employing a cohort of postgraduates in information science who will be assigned genres from Table 1 in Kim & Ross (2006c). They will retrieve one hundred PDF documents for each of the genres they have been assigned, and give a brief description of the source of the document and the reasons for including the document in their collection.

2. **Re-labelling Experiment:** we will anonymise the file names of the documents collected in the DRE and randomise the document sequence. This corpus will be presented to two new groups of labellers drawn from different backgrounds for re-classifying. They will not have access to the initial genre classification information.

The first experiment will create a pool of PDF files which have already been classified into genres by established organisations and users; this will serve as a reference point, and help us to index the performance on well-designed classification standards. The re-labelling experiment will enable us to compare the disagreement of the three classes of labellers over the same data set: this will help to determine the maximum level of accuracy at which the automated system can be expected to perform and determine which genres are better defined by looking at percentage of files in agreement within each genre.

The longer term aim, once a genre classifier or identifier with performance comparable to an average human labeller has been developed, will be to integrate the method with other tools which extract author, title, date, identifier, keywords, topic, language, summarisations and other compositional properties and objects (tables, links, figures) of files within a single genre. As we mentioned in Section 1, this will help to create the context necessary for understanding scientific data. The construction of a genre classification or detection tool, even without further extraction of metadata would be useful already in quickly and efficiently searching for and retrieving the scientific materials necessary for interpreting and carrying out scientific research.

6 ACKNOWLEDGEMENTS

This research is a part of The Digital Curation Centre's (DCC) research programme. The DCC is supported by a grant from the United Kingdom's Joint Information Systems Committee (JISC) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC) (grant GR/T07374/01) provides the support for the research programme. Additional support comes from the DELOS: Network of Excellence on Digital Libraries (G038-507618) funded under the European Commission's IST 6th Framework Programme. We also extend our thanks to Volker Heydegger at the Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI), University of Cologne, for his programming expertise; HKI participates in the DELOS Digital Preservation Cluster led by the University of Glasgow.

Note on website citations: All citations of websites were validated on 22 November 2006.

5 REFERENCES

Automatic Metadata Generation: http://www.cs.kuleuven.ac.be/hmdb/amg

Bagdanov, A. D. & Worring, M. (2001) Fine-Grained Document Genre Classification Using First Order Random Graphs. I*ntnl. Conf. Document Analysis and Recognition,* 79.

Barbu, E., Heroux, P., Adam, S. & Trupin, E. (2005) Clustering Document Images Using a Bag of Symbols Representation. *Intnl. Conference on Document Analysis and Recognition,* 1216–1220.

Bekkerman, R., McCallum, A. & Huang, G. (2004) Automatic Categorization of Email into Folders. Benchmark Experiments on Enron and SRI Corpora', *CIIR Tech. Report*, IR-418.

Biber, D. (1995) *Dimensions of Register Variation:a Cross-Linguistic Comparison.* Cambridge University.

Boese, E. S.(2005) Stereotyping the web: genre classification of web documents. *Master's thesis* Colorado State University.

Curran, J. & Clark, S. (2003) Investigating GIS and Smoothing for Maximum Entropy Taggers. *Proceedings, Aunnual Meeting European Chapter of the Assoc. of Computational Linguistics,* 91-98.

Digital Curation Centre: http://www.dcc.ac.uk

DC-dot, UKOLN Dublin Core metadata editor: http://www.ukoln.ac.uk/metadata/dcdot/

DELOS Network of Excellence on Digital Libraries: http://www.delos.info/

Engineering and Physical Sciences Research Council (EPSRC): http://www.epsrc.ac.uk/

Finn, A. & Kushmerick, N. (2006) Learning to Classify Documents According to Genre. *Journal of American Society for Information Science and Technology,* 57 (11), 1506-1518.

Giuffrida, G., Shek, E. & Yang, J. (2000) Knowledge-based Metadata Extraction from PostScript File. *5th ACM Intnl. Conf. Digital Libraries,* 77–84.

Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z. & Fox, E. A. (2000) Automatic Document Metadata Extraction using Support Vector Machines. *3rd ACM/IEEE-CS Conf. Digital libraries* 37–48.

Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI), University of Koeln: http://www.hki.uni-koeln.de/

Joint Information Systems Committee: http://www.jisc.ac.uk/

Karlgren, J. & Cutting, D. (1994) Recognizing Text Genres with Simple Metric using Discriminant Analysis. 1*5th Conf. Comp. Ling.* Vol 2 1071–1075.

Ke, S. W., Bowerman, C. & Oakes, M. (2006) PERC: A Personal Email Classifier. *28th European Conf. Information Retrieval (ECIR 2006),* 460–463.

Kessler, B., Nunberg, G. & Schuetze, H. (1997) Automatic Detection of Text Genre. *35th Ann. Meeting ACL* 32–38.

Kim, Y. (2004) Anaphora Resolution for Automatic Citation Linking. *Masters Thesis Speech and Language Processing,* University of Edinburgh.

Kim, Y. & Ross, S. (2006a) Genre Classification in Automated Ingest and Appraisal Metadata. J. Gonzalo et al. (eds.): *ECDL 2006*, LNCS 4172, 63–74.

Kim, Y. & Ross, S. (2006b) Automating Metadata Extraction: Genre Classification *Poster at the UK e-Science All Hands Meeting,* Nottingham, UK. http://www.allhands.org.uk/2006/proceedings/papers/663.pdf

Kim, Y. & Ross,S. (2006c) “The Naming of Cats”: Automated Genre Classification. *Proc. 2nd International Digital Curation Conference,* 21-22 November, 2006*,* Glasgow, UK.

Kim, Y. & Webber, B.(2006) Implicit Reference to Citations: A study of astronomy papers. *Preprint at ERPAePRINTS*, ID Code 115, http://eprints.erpanet.org

McCallum, A. (1998) Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering.

MetadataExtractor: http://pami.uwaterloo.ca/ (follow the link for Text Mining)

Noonberg, D., B. (2006) XPDF PDF document viewer. http://www.foolabs.com/xpdf/

PDF, Adobe Acrobat specification: http://partners.adobe.com/public/developer/pdf/index_reference.html

PDFTOHTML, PDF to HTML converter. http://pdftohtml.sourceforge.net/

PREMIS (PREservation Metadata: Implementation Strategy) Working Group: http://www.oclc.org/research/projects/pmwg/

Python: http://www.python.org

Python Imaging Library: http://www.pythonware.com/products/pil/

Rauber, A. & Müller-Kögler, A. (2001) Integrating Automatic Genre Analysis into Digital Libraries. *ACM/IEEE Joint Conf. Digital Libraries,* Roanoke, VA.

Riloff, E.,Wiebe, J., & Wilson, T. (2003) Learning Subjective Nouns using Extraction Pattern Bootstrapping. *7th CoNLL,* 25–32.

Santini, M. (2004a) A Shallow Approach To Syntactic Feature Extraction For Genre Classification. 7th Ann. Colloq. UK Special Interest Group for Comp. Ling.

Santini, M. (2004b) State-of-the-art on Automatic Genre Identification. Tech. Report ITRI-04-03 ITRI University of Brighton, UK.

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys,* Vol. 34 (2002) 1-47

Thoma,G. (2001) Automating the production of bibliographic records. *R&D report of the Communications Engineering Branch,* Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Witte, R., Krestel, R. & Bergler, S. (2005) ERSS 2005:Coreference-based Summarization Reloaded. *DUC2005 Document Understanding Workshop,* Canada

Witten, I., H. & Frank, E. (2005) *Data Mining: Practical machine Learning tools and techniques.* 2nd Edition, Morgan Kaufmann, San Francisco, USA.